PROJECTION OF RESULTS ON THE NAEP SCALE, USING DATA FROM THE NORTH CAROLINA END-OF-GRADE TESTING PROGRAM

David Thissen, Kathleen Billeaud, Lori Davis, *The University of North Carolina at Chapel Hill*
Eleanor Sanford, *North Carolina Department of Public Instruction*
Valerie S. L. Williams, *National Institute of Statistical Sciences*
Valerie S. L. Williams, *NISS*, P.O. Box 14162, Research Triangle Park NC 27709-4162

Key Words: National Assessment of Educational Progress (NAEP), Item Response Theory (IRT), Bootstrap

Introduction

The *Goals 2000: Educate America Act* requires an instrument to assess the effects on student performance of education reform, and to monitor progress with respect to consensual national achievement standards. One approach to fulfilling this requirement is the establishment of linkages between state testing programs and a common metric of growth and change, such as the scale used for the National Assessment of Educational Progress (NAEP). With such linkages, results from more frequently-administered state assessments could be translated into estimates of results that would have been obtained had the NAEP Trial State Assessment (TSA) been administered. This would reduce the reliance on a national testing program such as the TSA for purposes of tracking student achievement, and facilitate the comparability of student outcomes across assessment instruments, across different education programs, and across states or other jurisdictions.

Not only could linkages serve to estimate state-NAEP results, but they could also provide comparable measures at the level of the local school district, or possibly at the level of the school building—neither the NAEP nor the TSA sampling designs currently support valid inferences for students, schools, or even school districts. Although Selden makes "the case for district- and school-level results from NAEP" (in Glaser & Linn, 1992), he concludes that (p. 96):

> If linking became available and economically feasible, it could be expected that states would use it to maintain particular features of, and purposes for, their testing programs, while tying into NAEP and generating NAEP scores for schools and districts. It would appear to be in the interest of assessment for states to be encouraged and enabled to develop statewide systems which are distinctive and creative, while tying into a national assessment system that provides local schools and systems valuable data in a common national currency.

There are currently no examples available for states to follow in linking locally-constructed tests to the NAEP scale. Bloxom, Pashley, Nicewander, and Yan (1995) have described a linkage of scaled scores on the Armed Services Vocational Aptitude Battery with NAEP; however, both the data and the analytic procedures used in that effort are more complex than are common for state assessments, primarily because it involved a number of subscale scores and multiple imputations (Rubin, 1987) of examinee proficiency. The present investigation reports the procedures and results of one successful attempt at linking a statewide assessment program to the NAEP scale using projection methodology. This study provides a practical model and explicates a set of procedures that can be followed in linking related but disparate tests.

Development of the NC-NAEP linkage

The North Carolina Department of Public Instruction has developed a comprehensive academic testing program for grades 3 through 8, the End-of-Grade (EOG) tests, for assessing achievement in mathematics, reading, social studies, and science. The EOG score scale for mathematics, vertically equated to describe the performance of students across grades, ranges from about 100 to 200, with an eighth-grade average of approximately 169 in 1994. The NAEP mathematics scale ranges from 0 to 500, with an eighth-grade mean of approximately 262 for the nation in 1990 and 266 in 1992. In 1990, North Carolina eighth-graders averaged 250 on the NAEP TSA in mathematics, and in 1992, 258.

Mathematics proficiency, as measured by the NAEP exercises, is not identical to mathematics proficiency as measured by the EOG tests. Nevertheless, there is considerable overlap in the content frameworks, despite the fact that the two tests were built to different specifications. Projection makes use of the empirical relation between scores on tests that do not measure exactly the same thing to predict the distribution of scores on one test (e.g., NAEP) from the distribution of another test (e.g., a state assessment).

Data collection. Eighth-grade examinees were selected in a two-stage sampling design where the primary sampling unit is the school: 103 schools were drawn, and 99 participated. A target sample of 30

students was randomly selected in each of the schools; actual counts ranged from 21 to 33 participants. A total of 2824 students were tested. Because the numbers in the "Native American," "Hispanic," and "Asian/Pacific Islander" ethnic classification categories were inadequate for separate projections, two ethnic classifications reflecting relative educational advantagement were created for the projection analyses: BHN ("Black," "Hispanic," and "Native American" examinees) and WA ("White," "Asian/Pacific Islander," and "Other" examinees).

The NC-NAEP linkage test administered in February 1994 contained 78 items, including a short form of the EOG mathematics test for grade 8 (40 multiple-choice items) and two blocks of released 1992 NAEP mathematics items (38 items: 29 multiple-choice and 9 free-response). Coefficient alpha for the summed scores of the 38 NAEP items is $\alpha = 0.88$, and $\alpha = 0.82$ for the 40 EOG items. The reliability of the combined 78-item test is $\alpha = 0.91$.

The procedures used to project the NAEP scaled score distribution require IRT item parameters for the two blocks of NAEP items that were administered to the linkage sample. Published item parameter estimates from the NAEP TSA documentation are based on separate within-subscale item calibrations using unidimensional two- and three-parameter logistic item response models. For this study, the Educational Testing Service (ETS) provided estimates of *a*, *b*, and *c* parameters for each item from a unidimensional three-parameter logistic model with proficiency defined as the principal axis obtained in an item analysis of the entire 1990 and 1992 NAEP item pool.

Selection of a model for NAEP averages and standard deviations, conditional on EOG scores and background variables. For each student, a NAEP posterior distribution is obtained based on the individual response pattern, the population distribution, and the IRT parameter estimates provided by ETS. The prior, also provided by ETS, is a non-Gaussian histogram for the 1992 national NAEP, derived from an analysis of the 1990 and 1992 NAEP tests. Each examinee's posterior distribution is represented by a probability polygon defined by the relative likelihood of that examinee's response pattern at 37 equally-spaced values of proficiency, the *quadrature points*. These distributions were rescaled, so that the height at each quadrature point is a proportion of 1.0 and each individual's posterior sums to 1.0. The sum of these distributions, weighted by the sampling weights, is the sample estimate of the 1994 distribution.

The EOG summed scores are transformed to EOG scaled scores, and students are categorized into groups based on ethnic classification and EOG scaled score. By summing the weighted posteriors for each ethnic classification × EOG score combination, two distributions of NAEP scores for each EOG scaled score category are created.

The projection equations fit the posterior mean of each ethnic classification × EOG score category as the dependent variable; this is the mean of the posterior distribution created by summing all the individual posteriors for each examinee in an ethnic classification × EOG score category. The predictors are ethnic classification (dummy-coded BHN = 0 and WA = 1) and EOG score category, centered by subtracting 165, the mean EOG scaled score for the linkage sample. The standard deviations of the ethnic classification × EOG score category posteriors are predicted from EOG score category only. Weighted least squares regression analysis, in which the ethnic classification × EOG score category subgroupings are weighted by the number of students in each subgrouping, produced the parameter estimates shown in Table 1. Inclusion of the ethnic classification × EOG score category interaction did not contribute significantly to the prediction of the means of the posteriors. The means for all the score categories, and the two regression lines, are shown in Figure 1; the standard deviations are similarly shown.

Table 1. Parameter estimates from the weighted least squares regression model for projecting February NAEP results from February EOG results.

For the prediction of ethnic classification × EOG score category posterior means:

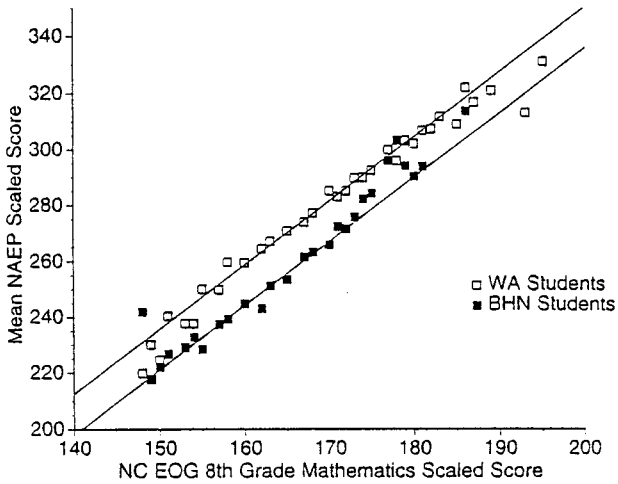
Variable	Coefficient (*se*)
Centercept	255.80 (0.91)
WA	14.11 (1.10)
EOGscore − 165	2.29 (0.06)

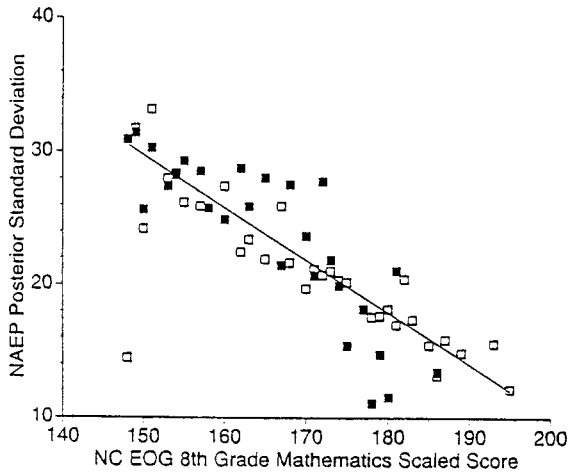
For the ethnic classification × EOG score category posterior standard deviations:

Variable	Coefficient (*se*)
Centercept	23.60 (0.32)
EOGscore − 165	−0.39 (0.03)

Bootstrap computation of standard errors for regression coefficients. The nested sampling design precludes inferences based on estimates of uncertainty calculated according to assumptions of simple random sampling. Standard errors for the regression coefficients were computed using a bootstrap procedure described by Sitter (1992a, 1992b). The bootstrap plan included finite population corrections at the first and second

Figure 1. Means for each ethnic classification × EOG score category, and the fitted regression lines (left), and the standard deviations for each ethnic classification × EOG score category, and the fitted regression line (right).





sampling stages, for school and for student-within-school. In practice, the finite population correction resamples n^* schools selected with replacement from the 99 schools, and m^* students selected with replacement from each school. According to Sitter (1992a):

$n^* = (n-1)/(1-f_1)$

where n is the number of clusters in the sample, N is the total number of clusters in the population, and $f_1 = n/N$, and

$m_i^* = (m_i-1)N/(1-f_{2i})n^*$

where m_i is the number of students in the ith sample cluster, M_i is the total number of students in the ith cluster, and $f_{2i} = m_i/M_i$. There are $N = 658$ schools with eighth grades in North Carolina, and $n = 99$ schools are represented in the NC-NAEP linkage sample, resulting in $n^* = 115$ schools to be resampled. The sizes of the eighth-grade classes, M_i, range from 29 to 496, and the size of the school-level samples, m_i, range from 21 to 33; the adjusted school-level sample sizes for the bootstrap, m_i^*, range from 24 to 3204, although the maximum was set to 300 to reduce computation.[1]

The bootstrap involves four steps:

Step 1 From the set of 99 schools, 115 schools are randomly selected with replacement; by chance, some schools are unrepresented, some duplicated, some triplicated, etc.

Step 2 From each school, m_i^* (between 24 and 300) students are randomly selected with replacement; again, some students are unrepresented, duplicated, etc.

Step 3 Using the data obtained in Step 2, the mean and standard deviation for each ethnic classification × EOG score category are calculated, and the projection equations computed to obtain the five regression coefficients.

Step 4 Steps 1 through 3 are repeated a total of 200 times, producing 200 estimates for each statistic.

To obtain the bootstrap estimate of a parameter, the mean of each set of 200 estimates is calculated, and the standard error of each of the statistics is the standard deviation computed from the sampling distribution. The bootstrap parameter estimates are not used, but the standard errors from the bootstrapped regression with finite population corrections appear in Table 1 with the weighted least squares regression coefficients.

Figure 2 shows the smoothed (Gaussian) posterior distributions of NAEP proficiency for three EOG summed scores, for BHN and WA examinees. The posteriors were approximated using a Gaussian distribution, with the mean obtained from the regression for the means, and the standard deviation obtained from the regression for the standard deviations.

Projection of February NAEP results from the May EOG administration. A second analysis projected the February NAEP results from the regular May administration of the EOG test. A total of 2313 students from the NC-NAEP linkage sample were matched with their May EOG scores; the average EOG increased about five points for this sample (to $\bar{X} = 169$). For future prediction of NAEP from the regular administration of the EOG, parameter estimates were again obtained using weighted least squares, and standard errors for the parameter estimates were computed using the bootstrap procedure, as described above. Table 2 contains these regression coefficients and the standard errors for

[1] For four of the 99 schools, m_i^* exceeded 300; the actual values are 315, 409, 996, and 3204.

Figure 2. Smoothed (Gaussian) posterior distributions of NAEP proficiency for three EOG scaled scores (155, 165, and 175), for examinees in the BHN (dashed lines) and WA (solid lines) ethnic classifications. Shown at right is the total NAEP posterior distribution, the weighted sum of the conditional distributions for all scores (140 to 204).

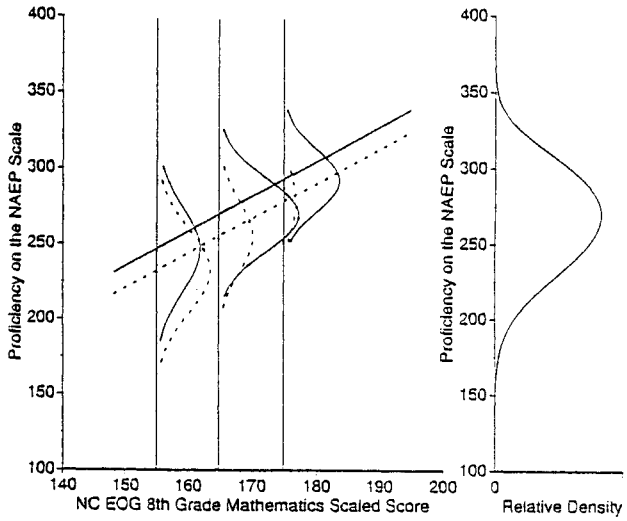


Table 2. Parameter estimates from the weighted least squares regression model for projecting February NAEP results from May EOG results.

For the prediction of ethnic classification × EOG score category posterior means:

Variable	Coefficient (*se*)
Centercept	259.96 (0.87)
WA	9.40 (1.08)
EOGscore − 169	2.25 (0.04)

For the ethnic classification × EOG score category posterior standard deviations:

Variable	Coefficient (*se*)
Centercept	21.12 (0.33)
EOGscore − 169	−0.30 (0.03)

predicting February's NAEP results from the May EOG test administration; the values differ very little from those in Table 1.

Computation of standard errors for the statistics derived from the projection. An empirical bootstrap procedure was used to compute the complete covariance matrices of the five regression parameters involved in the projections; the standard errors of the regression coefficients reported in Tables 1 and 2 are the square roots of the diagonal elements of those matrices. Simulation is used to compute estimates of the standard errors of the statistics derived from the projection, such as the projected percentiles. We simulate the effects that the use of different samples to develop the projection might have on the statistics derived from the projection.

To accomplish the simulation, it is assumed that the five regression coefficients are drawn from a multivariate normal distribution, with the mean equal to the estimates and the covariance matrix computed with the empirical bootstrap. Using as the five regression parameters random draws from that multivariate normal distribution, 200 projected values are obtained. The standard deviations of the derived statistics, such as the percentiles, computed over the 200 projected values, are reported as the standard errors of the derived statistics. Efron and Tibshirani (1993, p. 53) refer to such simulation for computing standard errors as the "parametric bootstrap," and describe the motivation for the procedure as a means "to provide answers in problems where no textbook formulae exist" (p. 55).

1994 State Results

The 1994 NAEP TSA results for North Carolina were obtained directly from the linkage sample. Subsequently, when the data from the statewide census administration of the EOG test became available, the projection equations summarized in Table 2 were developed, and the data from all 82,657 eighth-grade students were used to project (or, in this case, *postdict*) the February NAEP results. Comparison of the estimated proficiency distribution from the projection with that obtained directly from the NAEP administration showed that the two distributions correspond closely.

NAEP TSA results are most commonly reported as a set of quantiles of the distribution. Table 3 shows the values of the percentiles typically reported, as observed in the 1994 special administration of NC-NAEP to the linkage sample of 2824 students, and as projected from the population of 82,657. Six of the seven percentiles from the projection are within one standard error of the original sample values, and the seventh is well within two standard errors. It should be noted that the standard errors for the projected values are smaller than those computed with the original sample. These standard errors take into account only the sampling variation in the projection itself: Because the data from which the projection is done are population values, there is no sampling variation from that source. Measuring the population with the wrong test results in less sampling variation than making an inference to the population with data from the right test but using a smaller sample.

(There are, of course, both systematic and random sources of error that are not captured in sampling variation; those sources of error are not reflected in either set of standard errors.)

Table 3.
Observed and projected percentiles for the distribution of mathematics proficiency for North Carolina eighth-grade students (bootstrap standard errors are shown in parentheses).

Percentile	Observed	Predicted
5th	206 (2.0)	208 (1.4)
10th	220 (2.0)	221 (1.1)
25th	244 (1.7)	243 (0.8)
50th	267 (1.7)	268 (0.6)
75th	291 (1.3)	291 (0.5)
90th	308 (1.4)	310 (0.6)
95th	319 (1.3)	320 (0.6)

When the data from the 1995 administration of the EOG eighth-grade mathematics test become available, it will be possible to project from those data the state's 1995 NAEP TSA results.

Projection of results for school districts. Average NAEP scores were projected for North Carolina's 119 school districts. These estimates of district-level mathematics performance show a large amount of variability within the state of North Carolina, with school district averages ranging from 239 to 286. This maximum value represents mathematics performance comparable to the highest state NAEP averages. For example, in 1992, the eighth-graders in Iowa and North Dakota averaged 283. The lower value indicates poor student performance similar to that in states such as Mississippi ($\bar{X}$ = 246) and Louisiana ($\bar{X}$ = 249), or in the District of Columbia and Guam ($\bar{X}$ = 234).

Discussion

As Mislevy (1992) notes, projection methodology is "rather precarious" (p. 63), largely because it relies on the empirical relation between qualitatively different evidence about the proficiencies of individuals and groups. One source of statistical uncertainty is model misspecification: Either the IRT or the (linear) regression models could be incorrect, or the assumed population distribution could be erroneous. There is uncertainty due to the sampling error associated with the calibration sample, as well as the error in the projection sample. (However, in the latter case, the error associated with the projection sample may be negligible—for the North Carolina testing program, the entire population is tested.)

The standard errors reported for NAEP tests are jackknifed estimates of uncertainty; the NC-NAEP linkage produced standard errors for the regression coefficients by computing the standard deviations of the bootstrapped distributions. Longford (1995) found that jackknifed estimates are biased, possibly because they neglect the within-cluster variability. The bootstrap technique used here includes both within-cluster variability and finite sample corrections, but raises questions requiring a more precise specification of the sources of uncertainty that are to be included in the description of variation for statistics derived from NAEP and other such assessments.

Other potential issues of concern are:

Each NAEP mathematics item is associated with one of five subscales, with item parameter estimates based on separate within-subscale calibrations. For this projection study, however, a unidimensional item response model was used with the NAEP items.

ETS has developed the *plausible values* technology, using multiple imputations (Rubin, 1987), especially for analysis of NAEP data. For analytical purposes, an examinee is assigned five values for each subscale score, each randomly drawn from the examinee's posterior proficiency distribution. The NC-NAEP linkage analyses used pointwise representations of the examinee's posterior distributions. (In theory, this computational difference should have no effect on the results.)

The population distribution used in the NAEP TSA is conditioned on a large number of principal components of variables collected from an extensive background questionnaire administered with the NAEP cognitive tests. The choice of conditioning variables affects the size of the root mean squared error of all parameter estimates in a predictable manner, i.e., greater population variance will translate into larger error variance. In the NC-NAEP linkage, no conditioning background variables were used.

Motivational differences are cited by both Bloxom et al. (1995) and Ercikan (1993) as possible reasons for the failure of test linkages, and the importance of motivational factors should not be underestimated here.

Each of the above concerns should be considered challenges to the interpretation of the NC-NAEP projection results. They are important issues that remain to be resolved by further study.

Conclusions

Because of the great expense involved in expanding NAEP to provide scores below the state level, a network of state-NAEP linkages may provide a more feasible solution for NAEP score reporting at the school district

level. North Carolina has developed a student achievement testing program which also serves as one mechanism for school district accountability. The NC-NAEP linkage will not only permit the state to make district-level comparisons to national data, but it also allows comparisons of school district progress with respect to national trends and standards. NAEP linkages would also facilitate state- and district-level comparisons with international results.

References

Bloxom, B., Pashley, P., Nicewander, A., & Yan, D. (1995). Linking to a large-scale assessment: An empirical evaluation. *Journal of Educational and Behavioral Statistics, 20,* 1-26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Ercikan, K. (1993). Predicting NAEP. Unpublished manuscript. Monterey, CA: CTB Macmillan/ McGraw-Hill.

Glaser, R., & Linn, R. (Eds.). (1992). *Assessing student achievement in the states.* Stanford, CA: National Academy of Education.

Longford, N. L. (1995). *Model-based methods for analysis of data from the 1990 NAEP Trial State Assessment.* Research and Development Report. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1987). *Multiple imputation for non-response in surveys.* NY: Wiley.

Sitter, R. R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association, 87,* 755-765.

Sitter, R. R. (1992b). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics, 20,* 135-154.